Predikcija cene nekretnine na osnovu teksta oglasa, slika i geografske lokacije nekretnine

Mladen Vidović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
mladenvidovic@uns.ac.rs

Ivan Radosavljević
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
ivanradosavljevic@uns.ac.rs

Aleksandra Mitrović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
aleksandramitrovic@uns.ac.rs

***Apstrakt*—Objektivno određivanje adekvatne cene nekretnine predstavlja jedan od uslova za njenu prodaju, odnosno kupovinu. Pomoć pri određivanju prodajne cene vlasnici nekretnina često traže od agenata specijalizovanih za prodaju. Čak i njihove procene često ne budu objektivne. Neretko se i prilikom procene cene nekretnine zanemaruju i ostali uticaji, poput okoline. Formiranje modela za objektivnu procenu cene nekretnine bi bilo od značaja i potencijalnim kupcima, kao i prodavcima nekretnine. Kupci bi time bili sigurni da je zahtevana cena adekvatna, a prodavci bi bili sigurni da njihova cena nekretnine nije potcenjena. U ovom radu predložen je model za utvrđivanje adekvatne cene nekretnine koji pored tehničke specifikacije nekretnine upotrebljava i slike nekretnine, kao i podatke o kvalitetu okoline nekretnine pribavljene na osnovu geografske lokacije. Za prepoznavanje objekata na slikama upotrebljena je neuronska mreža koja je obučena na slikama dobijenim iz oglasa nekretnina. Objekti prepoznati od strane neuronske mreže su integrisani u početni skup podataka. Takođe, skup podataka je proširen i podacima dobijenim klasterovanjem nekretnina spram objekata u okolini. Osnovni skup podataka je podeljen na podskup koji sadrži podatke o koordinatama i podskup koji sadrži podatke o slikama. Svaki skup podataka je podeljen na obučavajući skup, na kom je vršena obuka, validacioni na kojem je vršena optimizacija modela i test na kom su evaluirani modeli. Za potrebe ovog rada su obučavani sledeći modeli: linearna regresija, linearne regresije sa lasso, ridge i elastic net regularizacijama, regresiono stablo, AdaBoost ansambl regresionih stabala i Gradient Boosted Trees. Najbolji model nad svim skupovima podataka je Gradient Boosted Trees.**

Ključne reči—nekretnine; slike; geografske lokacije; regresija

I. UVOD

Tržište nekretnina u velikim gradovima Srbije je veoma aktivno. Najaktivnije je u Beogradu i Novom Sadu zbog velikog broja stanovništva i brzog razvoja gradova. Cenu nekretnine određuju njihovi vlasnici i agenti za prodaju nekretnina. Procena cene nekretnine se vrši na osnovu osobina nekretnine, kao što su na primer: površina, stanje, opremljenost, lokacija itd. Njihove procene često nisu egzaktne i objektivne. Nerealne i previsoke cene i loše procene otežavaju prodaju nekretnina. Uvođenje objektivnog metoda za procenu cene nekretnine bi sa jedne strane pomoglo potencijalnim kupcima tako što bi im potvrdilo da je zahtevana cena prikladna, a prodavcima bi pomogla pri odabiru adekvatne cene.

U ovom radu će biti predstavljeno jedno rešenje za objektivno određivanje cene nekretnine. Rešenje će biti realizovano kao model za predikciju cene nekretnine obučen na osnovu podataka iz oglasa. Pored tekstualnog opisa oglasa, za obučavanje modela će se koristiti i slike nekretnine, kao i podaci o okolini dobavljeni na osnovu geografske lokacije nekretnine.

Izazovi pri realizaciji ovog rešenja su mnogobrojni. Tekstulani opisi oglasa su često nepotpuni ili netačni. Cene nekretnina su često postavljenje van razumnih opsega, ili se čak mogu pribaviti samo na zahtev. Nepouzdatnost podataka je posledica nedostatka bilo kakve validacije prilikom objavljivanja oglasa. Slike nekretnine iz oglasa su često lošeg kvaliteta. Pored toga, uz mnoge oglase nisu ni priložene slike nekretnine. Vodeni žig oznaka agencija na slikama takođe predstavlja dodatni izvor šuma. U oglasima često nisu zadate koordinate stana ili adresa na osnovu koje se mogu zaključiti koordinate, što onemogućava dobavljanje podataka o kvalitetu okoline nekretnine.

Detaljniji opis podataka, izazova i rešenja izložen je u ostatku rada. Naredno poglavlje se bavi srodnim istraživanjima na ovu temu. U trećem poglavlju će biti opisan skup podataka i način pripreme podataka za obučavanje i validaciju modela. Nakon toga, biće predstavljena metodologija koja je korišćena za rešavanje problema predviđanja cene nekretnine. Potom sledi prikaz rezultata. Na kraju će biti izveden zaključak ovog rada.

II. SRODNA ISTRAŽIVANJA

U radu [1], Ahmed i Moustafa predstavljaju rešenje za predikciju cene kuće na osnovu tekstualnih i vizualnih podataka, odnosno slika. Dodavanjem slika u skup podataka, ostvarili su bolje rezultate nego sa modelom obučenim bez slika. Konačan skup podataka sa kojim su radili je obuhvatao samo 535 kuća, što je dovelo do preprilagođavanju modela već nakon nekoliko koraka obučavanja.

U radu [2], Ottensman et al. uključuju lokaciju nekretnine u skup podataka, u vidu udaljenosti od poslovnog centra grada, kao i značajnih zona zaposlenja. Pokazalo se da lokacija nekretnine ima uticaj na cenu, iako je manji od uticaja karakteristika nekretnine. Ustanovili su da je cena nekretnine obrnuto srazmerna udaljenosti od značajnih zona. Pri određivanju kvaliteta lokacije nekretnine, vreme putovanja u periodima gustog saobraćaja do značajnih zona se pokazalo kao bolja mera od same udaljenosti do tih zona.

Goodman i Thibodeau u radu [3] pokazuju da lokacija nekretnine, odnosno ZIP kod, unutar većeg grada utiče na cenu nekretnine. Obrazložili su uticaj ZIP koda na cenu time što su nekretnine koje imaju isti ZIP kod geografski blizu, pa imaju slične objekte u okolini. Iz ovoga su autori zaključili da objekti u okolini nekretnine utiču na njenu cenu.

Kong et al. su u radu [4] prikazali uticaj javnih objekata i posebno zelenih površina u okolini nekretnine na njenu cenu. Za dobavljanje podataka o tim objektima, radius od 500 metara se pokazao kao najbolji. Dokazali su da na cenu nekretnine najviše pozitivno utiču univerziteti i parkovi. Međutim ukoliko se u okolini nekretnine nalazi previliki broj ovih objekata, njihov uticaj na cenu nekretnine postaje negativan, jer se pokazalo da ljudi preferiraju da žive u mirnijim, slabije naseljenim delovima grada.

III. OPIS SKUPA PODATAKA

Inicijalni skup podataka je formiran na osnovu podataka prikupljenih sa veb stranice za oglašavanje nekretnina nekretnine.rs [5]. Pri prikupljanju podataka izabrani su oglasi za nekretnine sa područja Beograda i Novog Sada. Ovo je urađeno zato što je tržište nekretnina najaktivnije u tim gradovima. Opseg ovog rada je ograničen na predikciju prodajne cene nekretnine, pa su odabrani samo oglasi za prodaju nekretnina. Od tipova nekretnina, odabrani su samo stanovi, jer se najveći broj oglasa odnosi na stanove. Ovako pribavljen skup podataka se sastojao od podataka za 95942 sirova oglasa. Za svaki oglas su prikupljene i fotografije nekretnine.

Svaki oglas sastoji se iz tehničke specifikacije u tekstualnom formatu i slobodnog opisa oglasa. Tehničku specifikaciju čine sledeći atributi:

- cena nekretnine izražena u evrima, koja za potrebe ovog rada predstavlja ciljni atribut,
- jedinstveni identifikator oglasa,
- površina objekta izražena u m^2,
- tip oglašavača koji može biti vlasnik, zastupnik, investitor ili agencija,
- datum objavljivanja oglasa,
- datum poslednjeg ažuriranja oglasa,
- adresa predstavljena nazivom ulice, brojem zgrade, delom grada, nazivom grada i državom, od kojih broj zgrade i deo grada ne moraju biti navedeni,
- podatak o tome da li je stan uknjižen,
- broj soba,
- broj kupatila,
- broj sprata na kojem se stan nalazi,
- ukupan broj spratova zgrade,
- tehnička opremljenost koja može biti klima uređaj, telefonski priključak, dostupnost interneta, kablovska televizija, interfon, video nadzor,
- dostupnost pomagala u koja spadaju lift i rampa za invalidska kolica,
- tip stana koji može biti standardni, dupleks, penthaus, garsonjera, potkrovlje, nadogradnja, salonac, dvorišni stan, nisko prizemlje, visoko prizemlje, prizemlje, stan u kući, suteren, pri čemu se za jedan stan može navesti više tipova,
- stanje objekta koje može biti standardna, novogradnja, starogradnja i stan u izgradnji,
- prateće površine u koje spadaju terasa, lođa, balkon, francuski balkon, bašta, bazen, ostava, parking, garaža i podrum,
- vrsta grejanja koja može biti centralno grejanje, daljinsko i etažno,
- vrste goriva za grejna tela, u koje spadaju lož ulje, gas, čvrsta goriva i struja,
- dostupnost gradskog prevoza pod čime se podrazumevaju autobus, tramvaj, voz i trolejbus,
- geografske koordinate nekretnine,
- spisak URL-ova dostupnih fotografija.

Eksplorativna analiza dobavljenih podataka je pokazala da neki oglasi nemaju navedenu cenu, odnosno navedeno je da se cena za njih može dobiti tek po zahtevu. Ovi oglasi su izbačeni iz skupa podataka. Takođe su uklonjeni oglasi za koje nije bilo navedeno na kojem spratu se nalaze. Odluka da se uklone oglasi bez podatka o spratu na kojem se stan nalazi donesena je zbog postojanja dovoljno jakog uticaja sprata na cenu stana [6], kao i zbog dovoljno velikog broja oglasa koji su imali podatak o spratu. Oglasi sa cenom ili površinom van razumnog opsega su takođe uklonjeni. Kao donja granica za cenu izabrano je 10000 evra, a za gornju je odabrano 1000000 evra. Odabrana donja granica za površinu je $10m^2$, a gornja $300m^2$. Daljom analizom utvrđeno je da za veliki broj stanova koji imaju samo jednu sobu ili jedno kupatilo u oglasu nije naveden broj soba, odnosno kupatila. Zbog toga je doneta odluka da se, ukoliko broj soba ili kupatila nije naveden, za broj soba, odnosno kupatila, kao podrazumevana vrednost postavi 1. Nakon prethodno navedenog prečišćavanja skupa podataka, preostalo je 61942 oglasa. Broj oglasa koji su premašili gornju granicu za cenu i površinu bio je 251. Oglasa koji su bili ispod donje granice za cenu i površinu je ukupno bilo 473. Bez cene je bilo 535 oglasa. Ukupan broj oglasa koji su izbačeni jer za njih nije naveden sprat na kojem se nalaze je iznosio 32742 oglasa.

Podaci iz prethodno navedenog skupa su ponovo prečišćeni spram prisutnosti podataka o koordinatama nekretnine. Prečišćavanjem je dobijen poseban skup od 19869 oglasa. Ovaj skup podataka se koristio za dobavljanje podataka o okolini nekretnine radi dalje procene njenog kvaliteta. Kvalitet okoline se određuje na osnovu dostupnih značajnih objekata i površina u njoj. Ovi podaci se mogu podeliti na sledeće kategorije:

- javni objekti poput škola, restorana, bolnica itd.,

- turistički objekti,
- zone sa specijalnom namenom, poput industrijske zone ili deponije,
- mesta za odmor i opuštanje.

Prikupljanje podataka o okolinama je urađeno korišćenjem Overpass API-a koji omogućuje dobavljanje geografskih podataka u proizvoljnom radijusu za proizvoljno zadate geografske koordinate [7]. Podaci su dobavljeni za četiri radijusa u okolini svake nekretnine. Prvo su formirana tri radijusa u opsezima: 0-100m, 100-500m i 500-1000m. Analizom uticaja dobijenih značajnih objekata i površina u datim radijusima na cenu, uočeno je da srednji i najveći radijus nisu imali značajan uticaj na cenu nekretnine. Detaljnijom analizom utvrđeno je da radijus 0-150m ima dovoljno značajan uticaj na cene nekretnina, pa je stoga odlučeno da se za potrebe ovog rada koriste podaci o objektima u tom radijusu. Eksplorativna analiza podataka o zastupljenosti objekata u prethodno navedenom radijusu je pokazala da se neki tipovi objekata uopšte ne pojavljuju u skupu podataka, iz tog razloga ovi tipovi su uklonjeni iz konačnog skupa podataka.

Na osnovu početnog skupa od 95942 oglasa izdvojeno je 662396 URL-ova za dobavljanje slika koje će činiti novi skup podataka. Uspešno je dobavljeno 571771 slika, od kojih je usled grešaka pri prikupljanju izgubljeno 138 slika. Ovaj skup podataka je upotrebljen za obučavanje neuronske mreže za prepoznavanje opremljenosti nekretnine. Ovo obuhvata već postojeće objekte iz specifikacije oglasa, poput grejanja i tehničke opremljenosti, ali i dodatne objekte poput nameštaja, sanitarija u kupatilu, kućnih aparata i sličnih objekata koji mogu da utiču na cenu nekretnine. Prepoznati objekti se koriste za verifikaciju datog opisa oglasa, kao i njegovo dopunjavanje ovim dodatnim objektima. Da bi se mogla obučiti neuronska mreža, bilo je neophodno prvo obeležiti značajne objekte na slikama. Obeležavanje je rađeno ručno koristeći LabelImg alat. Objekti koji su bili od značaja i koji predstavljaju labele su:

- tuš kabina,
- kada,
- WC šolja,
- umivaonik,
- veš mašina
- bojler,
- električni šporet,
- šporet na drva,
- običan ili kombinovani frižider,
- zamrzivač sandučar,
- sudopera,
- sudo-mašina,
- trpezarijski sto sa stolicama,
- radni sto,
- krevet, kauč, ugaona garnitura,
- ormar gde je uračunat gardarober i regal,
- TV,
- interfon,
- klima,
- radijator centralnog grejanja,
- električni radijator,
- kvarcna peć,
- TA peć.

Ukupno je obeleženo 7998 slika, od čega je 25% odvojeno za test skup. Prilikom analize zastupljenosti datih objekata u stanovima utvrđeno je da ne postoje primeri sa zamrzivačem sandučarom, kvarcnom peći, električnim radijatorom i šporetom na drva. Zbog toga su ove labele uklonjene. Na Slika 1 dat je prikaz zastupljenosti objekata pre uklanjanja pomenutih klasa.

Za oglase koji imaju i navedene koordinate i priložene slike, kreiran je poseban skup podataka, koji sadrži 9862 oglasa. U ovaj skup podataka su uključeni i podaci o okolini, i podaci dobijeni prepoznavanjem objekata na slikama.

U svim prethodno pomenutim skupovima podataka podaci iz oglasa su bili u obliku teksta. Za dalje procesiranje bilo je neophodno pretvoriti date podatke u numeričke. Pretvaranje je odrađeno tehnikama one hot encoding i dummy coding.

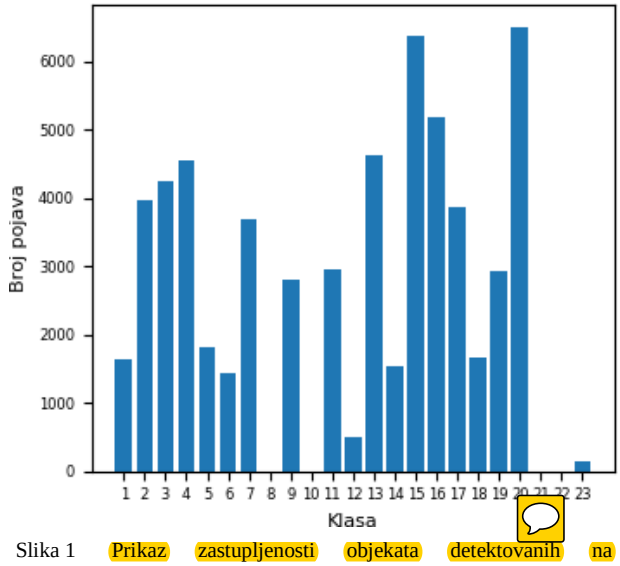
Konačno, tri dobijena skupa podataka podeljena su na podskupove za test, 20% podataka, validacione podskupove 20% od preostalih 80%, dok je ostatak podataka ostavljen za obučavanje modela. Za svaki od ovih skupova podataka izračunata je minimalna, maksimalna i srednja vrednost cene nekretnine. U osnovnom skupu podataka minimalna cena iznosila je 10000 evra, maksimalna cena 1000000 evra a srednja vrednost cene je iznosila 61870 evra. Cene u skupu podataka sa ocenom kvaliteta okoline kretale su se od 11000 evra do 950000, srednja vrednost iznosila je 64890 evra. Isti raspon cena imao je i skup sa podacima o kvalitetu okoline i detektovanim objektima u nekretnini dok je srednja vrednost za ovaj skup iznosila 66000 evra. Na Slika 2 su prikazani rasponi i raspodela cena u sva tri skupa podataka. Za pomenute skupove podataka izračunate su matrice korelacije kako bi se ustanovilo da li su neki od atributa u jakoj korelaciji. Jaka korelacija između atributa je indikator da jedan od iskoreliranih atributa treba da bude uklonjen. Na Slika 3 dat je prikaz dobijenih matrica korelacije. Analizom ovih matrica ispostavilo se da nema atributa koji su u dovoljno jakoj korelaciji da bi bili uklonjeni.

IV. METODOLOGIJA

Metodologija ovog rešenja može se podeliti u tri celine: analiza i integracija podataka o okolinama, obučavanje neuronske mreže za prepoznavanje objekata u stanu i izbor i optimizacija modela za predviđanje cene nekretnine.

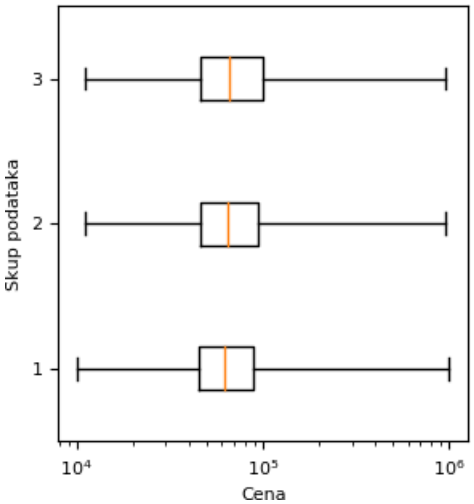
A. Analiza i integracija podataka o okolinama

Za integrisanje podataka o okolini nekretnine u prediktivne modele, nekretnine su grupisane prema sličnosti njihovih okolina. Ovo je postignuto primenom k-means klasterovanja. Broj klastera je određen empirijski, analizom objekata u okolini centroida klastera, kao i raspodelom cene po klasterima. Najinterpretabilniji je bio model sa dva klastera. Nekretnine su

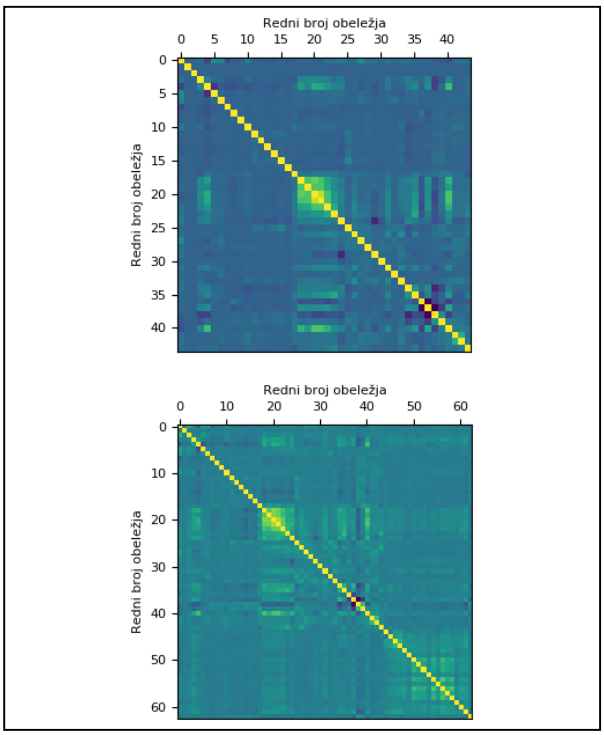


Slika 1 Prikaz zastupljenosti objekata detektovanih na fotografijama oglasa.

jasno podeljene na one koje u svojoj okolini nemaju mnogo objekata, i nekretnine okružene velikim brojem objekata. Klasterovanje je izvršeno i sa tri, četiri, pet i šest klastera. Sa porastom broja klastera, podaci su postajali manje interpretabilni, a razlike u cenama su postajale manje. Klaster sa malim brojem objekata se sa porastom broja klastera malo menjao, dok su se nekretnine sa većim brojem objekata grupisale prema tipu objekata, ali se njihova cena nije značajno razlikovala. Nakon klasterovanja, nekretnine su ponovo spojene sa podacima o njihovoj ceni. Pokazalo se da je medijan cena klastera sa malim brojem objekata u okolini manji od medijana cena drugog klastera za 15000 evra. Ovo predstavlja značajnu razliku u cenama, što potvrđuje da cena raste srazmerno porastu broja objekata u okolini nekretnine. Izuzetak ovome predstavljaju autobuske stanice, koje su dominantno zastupljene u klasteru sa nižom cenom, što potvrđuje pretpostavku da nekretnine u mirnijim krajevima imaju veću



Slika 3 Prikaz raspona i raspodele cena nekretnina u skupovima podataka. 1) Osnovni skup podatka, 2) skup podataka sa ocenama okoline, 3) skup podataka sa ocenama nekretnine i detektovanim objektima.



Slika 2 Matrice korelacija za osnovni skup podataka i skup podataka proširen ocenom kvaliteta nekretnine i detektovanim objektima sa fotografija nekretnine.

vrednost [4]. Model sa dva klastera je spojen sa svim ostalim podacima iz oglasa, i dobijeni skup podataka je kasnije upotrebljen za predikciju cena.

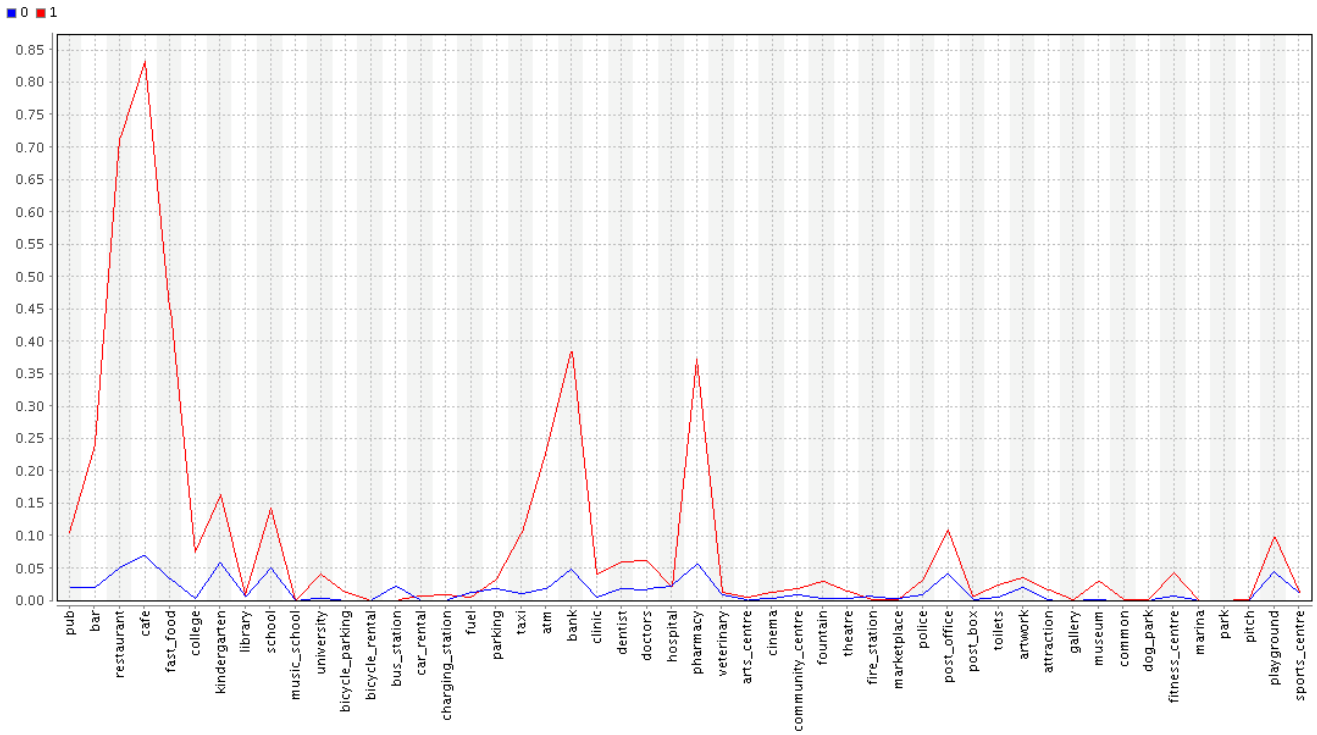
Na slici Slika 4 su prikazani centri klastera, odnosno zastupljenost objekata u okolinama karakterističnim za nekretnine u klasterima. Dok je na Slika 5 prikazana razlika srednjih vrednosti cena između klastera.

B. Prepoznavanje objekata u stanu

Za prepoznavanje objekata na fotografijama stanova upotrebljena je neuronska mreža. Pre obučavanja, urađena je augmentacija skupa podataka, radi izbegavanja preprilagođavanja mreže. Početno primenjene metode augmentacije bile su:

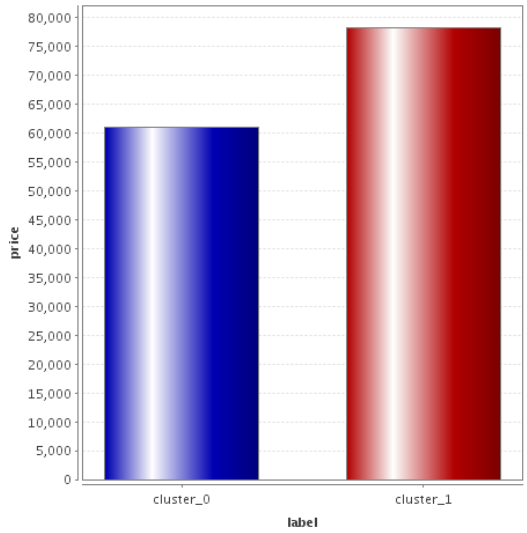
- horizontalno obrtanje slike,
- izmena osvetljenja slike,
- izmena kontrasta slike,
- izmena zasićenja boja slike.

Međutim, augmentacija izmenom kontrasta, osvetljenja i zasićenja boja slika nije primenjena u konačnoj augmentaciji skupa podataka, zato što je imala negativan uticaj na performanse mreže i brzinu obučavanja. Na ulaz neuronske mreže su dostavljane slike iz oglasa skalirane tako da im maksimalna visina i širina budu 800 piksela, a minimalna 600 piksela. Za obučavanje preuzeta je konvolutivna neuronska mreža, prethodno obučena na COCO skupu podataka [8], koji sadrži svakodnevne objekte slične onim koje je potrebno prepoznati u stanovima. Arhitektura mreže je Faster R-CNN ResNet101 [9]. Mreža je dodatno obučena nad prethodno



Slika 4 Prikaz zastupljenosti objekata u centrima klastera. Crvenom linijom je prikazan klaster u kojem su najviše zastupljene skupe nekretnine. Plavom linijom prikazan je klaster sa jeftinijim nekretninama.

prikupljenim slikama iz oglasa tako da prepoznaje 23 klase. Obučavanje je izvršeno u 50000 iteracija. Nakon obučavanja dobijena neuronska mreža je primenjena za detekciju objekata na slikama nekretnina iz oglasa koji su posedovali podatke o geografskoj lokaciji. Podaci o detektovanim objektima su potom spojeni sa podacima iz oglasa i na taj način je formiran novi skup podataka sačinjen od 9862 oglasa. Kasnije je ovaj skup podataka upotrebljen za predikciju cene nekretnine.



Slika 5 Srednje vrednosti cena nekretnina u klasterima.

C. *Predviđanje cene nekretnine*

Za potrebe predikcije cene nekretnine, odabrano je nekoliko regresionih modela:

- linearna regresija, kao osnova za poređenje sa drugim modelima,
- linearna regresija sa lasso regularizacijom,
- linearna regresija sa ridge regularizacijom,
- linearna regresija sa elastic net regularizacijom,
- regresiono stablo,
- ansambl regresionih stabala u AdaBoost konfiguraciji,
- ansambl regresionih stabala u Gradient Boosted Trees modelu.

Svaki regresioni model je primenjen na svakom podskupu podataka. Za svaki od ovih modela je urađena optimizacija parametara nad validacionim skupovima podataka. Mera korišćena za evaluaciju performansi modela je R^2.

Prvo je izvršena optimizacija parametara za lasso, ridge, i elastic net regresiju. Nakon optimizacije nije bilo značajnog poboljšanja performansi u odnosu na linearnu regresiju bez regularizacije. Zbog toga su odbačeni ovi modeli iz daljeg razmatranja.

Model regresionog stabla je optimizovan u odnosu na dubinu stabla. Za osnovni skup podataka, bez podataka o kvalitetu okoline ili slika, optimalna dubina stabla je 7. Za skup podataka oglasa koji imaju koordinate dobijena optimalna dubina iznosila je 4 pre dodavanja podataka o klasterima, a 5 nakon dodavanja klastera. Za skup podataka koji imaju i koordinate i slike, optimalna dubina je bila 4 i pre i posle dodavanja klastera i detektovanih objekata.

Model regresionih stabala u AdaBoost ansamblu je optimizovan prema dubini regresionog stabla i broja estimatora u ansamblu. Za osnovni skup podataka, optimalna dubina stabla je 14, a broj estimatora 80. Za skup podataka sa koordinatama je optimalna dubina stabla 13, a broj estimatora 40 pre dodavanja klastera, a dubina 11 i 60 estimatora nakon dodavanja klastera. Optimalna dubina stabla za skup podataka sa koordinatama i slikama, pre dodavanja klastera i detektovanih objekata je 12, a broj estimatora 160. Optimalna dubina stabla nakon dodavanja detektovanih objekata je 9, sa 50 estimatora, a nakon dodavanja i klastera, optimalna dubina je 13 sa 50 estimatora.

Gradient Boosted Tree regresioni model je takođe optimizovan prema dubini regresionog stabla i broja estimatora. Za osnovni skup podataka, optimalna dubina stabla je 12, a broj estimatora 70. Optimalna dubina stabla za skup podataka sa koordinatama je 7 sa 80 estimatora pre dodavanja podatka o klasterima, a 11 sa 70 estimatora nakon dodavanja klastera. Za skup podataka sa koordinatama i slikama, pre dodavanja klastera, optimalna dubina regresionog stabla je 7, a broj estimatora 80. Nakon dodavanja slika, optimalna dubina stabla je bila 11 sa 60 estimatora, a nakon dodavanja i klastera, optimalna dubina je bila 8 sa 70 estimatora.

U TABELA I su prikazane performanse optimizovanih regresionih modela na validacionom skupu. U redovima tabele su navedeni skupovi podataka, a u kolonama regresioni modeli.

V. REZULTATI

Prikaz rezultata će biti podeljen u dve celine. Prva celina se odnosi na postignute rezultate neuronske mreže pri prepoznavanju objekata na slikama iz test skupa. Drugi deo se odnosi na rezultate evaluacije regresionih modela na test skupu podataka. Na kraju svake celine sledi analiza grešaka modela.

A. *Neuronska mreža*

Obučena neuronska mreža je testirana na skupu od 1998 slika koje nisu bile uključene u obučavajući skup. Testiranje mreže je izvršeno koristeći evaluator unutar TensorFlow object detection API-a [10]. Pri evaluiranju, korišćene su dve metrike. Prva je mean average precision, koja predstavlja prosečnu preciznost za sve klase.

TABELA I. Rezultati na validacionom skupu nakon optimizacije modela.

Skupovi podataka	**Regresioni modeli**			
	Linearna regresija	***Regresiono stablo***	***AdaBoost***	***Gradient Boosted Trees***
Osnovni	0.694	0.703	0.717	0.77
Geo bez klastera	0.738	0.758	0.776	0.818
Geo sa klasterima	0.745	0.765	0.789	0.832
Geo-Img bez klastera i slika	0.679	0.689	0.725	0.758
Geo-Img bez klastera	0.684	0.689	0.729	0.759
Geo-Img	0.695	0.698	0.76	0.771

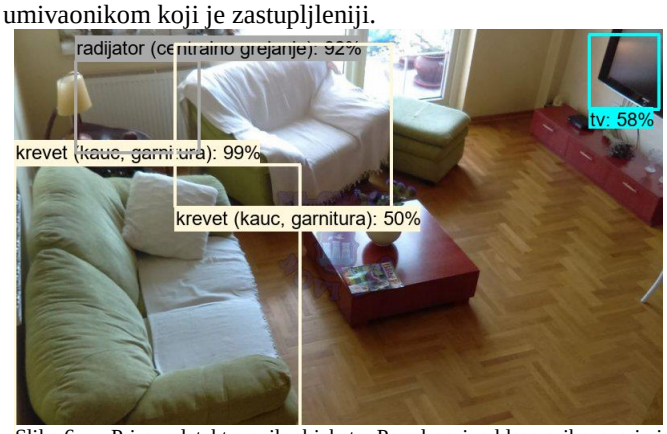
Druga upotrebljena metrika je average precision, što predstavlja preciznost za svaku klasu pojedinačno.

Vrednosti preciznosti za svaku klasu su:

- tuš kabina - 0.574
- kada - 0.739
- WC šolja - 0.79
- umivaonik - 0.77
- veš mašina - 0.707
- bojler - 0.703
- električni šporet - 0.735
- šporet na drva - *
- običan ili kombinovani frižider - 0.697
- zamrzivač sandučar - 0
- sudopera - 0.487
- sudo mašina - 0.589
- trpezarijski sto sa stolicama - 0.898
- radni sto - 0.591
- krevet, kauč, ugaona garnitura - 0.858
- ormar - 0.59
- TV - 0.788
- interfon - 0.476
- klima - 0.842
- radijator centralnog grejanja - 0.834
- električni radijator - 0.062
- kvarcna peć - 0
- termoakumulaciona peć - 0.333

Klase označene sa * su klase koje se nisu pojavile u test skupu. Loša preciznost za neke klase se može obrazložiti malim brojem primera za te klase u obučavajućem skupu. Globalna prosečna preciznost za sve klase je 0.594. Primer detektovanih objekata sa fotografije nekretnine dat je na Slika 6.

Za podatke koji su bili adekvatno zastupljeni u obučavajućem skupu, najviše grešaka je bilo vezano za uimvaonike, zato što su slični ostalim sanitarijama, i nekad su pomešani sa kadom ili bideom. Greške pri prepoznavanju ormara su nastale jer su dvokrilna vrata prepoznavana kao ormari. Električni radijatori su mnogo slični radijatorima centralnog grejanja, koji su ipak zastupljeniji u skupu podataka. Greške pri prepoznavanju sudopera se mogu objasniti velikom varijansom u izledu sudopere, kao i njenom sličnošću sa umivaonikom koji je zastupljeniji.



Slika 6 Primer detektovanih objekata. Pored naziva klasa prikazana je i sigurnost predikcije za datu klasu.

B. *Regresija*

Regresioni modeli su testirani na test skupovima izdvojenih iz skupova podataka koji su objašnjeni u trećem poglavlju ovog rada. Test skupovi su odvojeni na početku i nad njima nije vršena nikakva izmena ili optimizacija, odnosno modeli nisu obučavani nad test skupovima. Evaluacija je izvršena koristeći parametre dobijene u koraku optimizacije modela. Rezultati evaluacije su prikazani u TABELA II.

Na osnovu rezultata se može zaključiti da, u opštem slučaju, dodavanje podataka o okolini nekretnine značajno utiče na performanse modela. Jedini izuzetak za to je regresiono stablo, koje se pokazalo kao previše jednostavan model za primenu na ovom skupu podataka, dok dodavanje slika nije pokazalo značajna poboljšanja u performansama. Razlog za to je što je skup podataka sa obeleženim slikama mnogo manji u odnosu na ostale skupove podataka i uz to sadrži dobar deo podataka sa velikom varijansom cene koja nije dovoljno dobro objašnjiva opremljenošću nekretnine.

Gradient Boosted Trees se pokazao kao model sa najboljim preformansama za svaki skup podataka. Najbolju R^2 vrednost je imao za skup podataka sa koordinatama nakon dodavanja podataka o objektima iz okoline. Dodavanje podataka o objektima detektovanim na slikama je negativno uticalo na performanse modela. Slika 7 prikazuje performanse modela na test skupu.

Sa Slika 7 se vidi da model najviše greši na primerima sa velikom površinom, preko 175 m^2, zato što je kod njih najveća varijansa u ceni. Skuplji stanovi su često novije gradnje pa nisu još opremljeni, tako da njihova specifikacija i slike ne mogu da opravdaju visoku cenu. Takođe ovaj skup uključuje stanove koji su još u izgradnji, pa nemaju ispravne specifikacije. Na cenu nekretnine utiču i neki faktori koje modeli predstavljeni ovom radu ne uzimaju u obzir, poput udaljenosti od centra grada i drugih značajnih delova grada, stope kriminala i drugih socio-ekonomskih faktora.

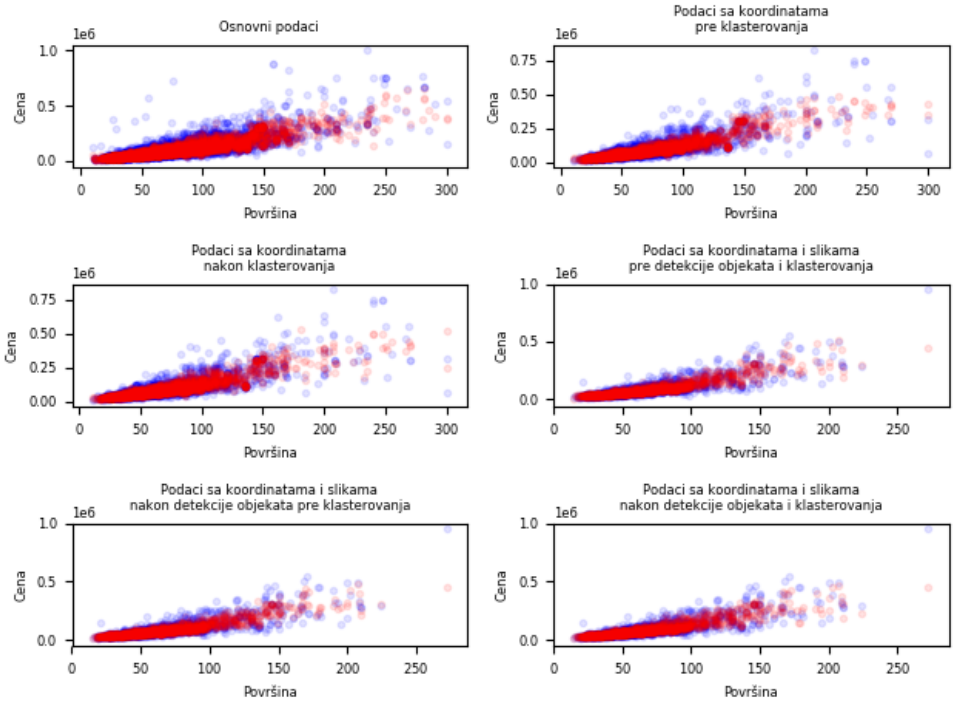
TABELA II. Rezultati na test skupu

Skupovi podataka	**Regresioni modeli**			
	Linearna regresija	***Regresiono stablo***	***AdaBoost***	***Gradient Boosted Trees***
Osnovni	0.69	0.71	0.741	0.757
Geo bez klastera	0.708	0.669	0.707	0.782
Geo sa klasterima	0.716	0.686	0.714	0.805
Geo-Img bez klastera i slika	0.73	0.688	0.712	0.778
Geo-Img bez klastera	0.734	0.688	0.748	0.772
Geo-Img	0.739	0.688	0.776	0.782

VI. ZAKLJUČAK

U ovom radu je prikazan model za utvrđivanje adekvatne cene nekretnine koji, pored tehničke specifikacije, uzima u obzir i podatke o okolini nekretnine, kao i fotografije nekretnine. Ovi podaci su prikupljeni iz oglasa preuzetih sa veb stranice nekretnine.rs.

Kvalitet okoline nekretnine je utvrđen analizom klastera nekretnina, dobijenih klasterovanjem objekata iz njihove



Slika 7 Predikcije Gradient Boosted Trees modela nad test skupom.

okoline. Utvrđeno je da je optimalan broj klastera dva, što deli nekretnine na one koji imaju mnogo objekata u svojoj okolini i one koji nemaju. Za nekretnine koje imaju mnogo objekata u svojoj okolini ispostavilo se da imaju i veću prosečnu cenu.

Fotografije nekretnina su upotrebljene za određivanje opremljenosti nekretnine. Utvrđivanje opremljenosti izvršeno je pomoću neuronske mreže koja je obučena na prethodno obeleženim slikama iz prikupljenih oglasa. Mreža je obučena da prepoznaje 23 klase u koje spadaju sanitarije, kuhinjska oprema i drugi objekti koji potencijalno utiču na cenu nekretnine. Prosečna preciznost postignuta u prepoznavanju ovih objekata iznosila je 0.594, što ne oslikava prave performanse jer je zastupljenost pojedinih klasa bila suviše mala.

Za predikciju cene nekretnine, obučeno je više regresionih modela, i izvršenja je optimizacija njihovih parametara. Upoređene su performanse svakog modela, nad osnovim skupom podataka, skupom podataka sa ocenom kvaliteta okoline i nad skupom podataka sa prepoznatom tehničkom opremljenošću i ocenom kvaliteta okoline. Mera izabrana za evaluaciju modela je R^2. Utvrđeno je da je najbolji model Gradient Boosted Trees, koji je postigao vrednosti R^2 od 0.757 na osnovnom skupu podataka, 0.805 na skupu podataka sa ocenom okoline i 0.782 na skupu podataka sa prepoznatom tehničkom opremljenošću i ocenom kvaliteta okoline.

Daljom analizom grešaka je ustanovljeno da model greši najviše kod primera sa velikom površinom, novijom gradnjom i nepotpunim specifikacijama, kao i nekretninama koje su još u izgradnji.

Planovi za dalji razvoj ovog rešenja obuhvataju:

- dobavljanje većeg broja obeleženih slika za obučavanje neuronske mreže u cilju poboljšavanja njenih performansi,
- formiranje ocene opremljenosti nekretnine klasterovanjem prema objektima detektovanim na slikama,
- proširenje skupa podataka sa podacima o udaljenosti nekretnina od centra grada i drugih značajnih objekata,
- izdvajanje korisnih informacija iz slobodnih opisa oglasa i njihova integracija u skup podataka,
- dobavljanje podataka o stopi kriminala kao i drugih socio-ekonomskih faktora i njihova integracija u skup podataka.

LITERATURA

[1] Eman H Ahmed and Mohamed Moustafa, "House Price Estimation from Visual and Textual Features," 2016.

[2] J. Ottensmann, S. Payton, and J. Man, "Urban Location and Housing Prices within a Hedonic Model," *J. Reg. Anal. Policy*, vol. 38, Jan. 2008.

[3] A. C. Goodman and T. G. Thibodeau, "Housing market segmentation and hedonic prediction accuracy," *J. Hous. Econ.*, vol. 12, no. 3, pp. 181–201, Sep. 2003.

[4] F. Kong, H. Yin, and N. Nakagoshi, "Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China," *Landsc. Urban Plan.*, vol. 79, no. 3, pp. 240–252, Mar. 2007.

[5] "Nekretnine – Najveći oglasnik za nekretnine u Srbiji." [Online]. Available: https://www.nekretnine.rs/. [Accessed: 15-Apr-2018].

[6] S. Conroy, A. Narwold, and J. Sandy, "The value of a floor: valuing floor level in high-rise condominiums in San Diego," *Int. J. Hous. Mark. Anal.*, vol. 6, no. 2, pp. 197–208, May 2013.

[7] "Overpass API — OpenStreetMap Wiki." [Online]. Available: https://wiki.openstreetmap.org/wiki/Overpass_API#Books. [Accessed: 15-Apr-2018].

[8] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," *ArXiv14050312 Cs*, May 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2016, pp. 770–778.

[10] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," Nov. 2016.